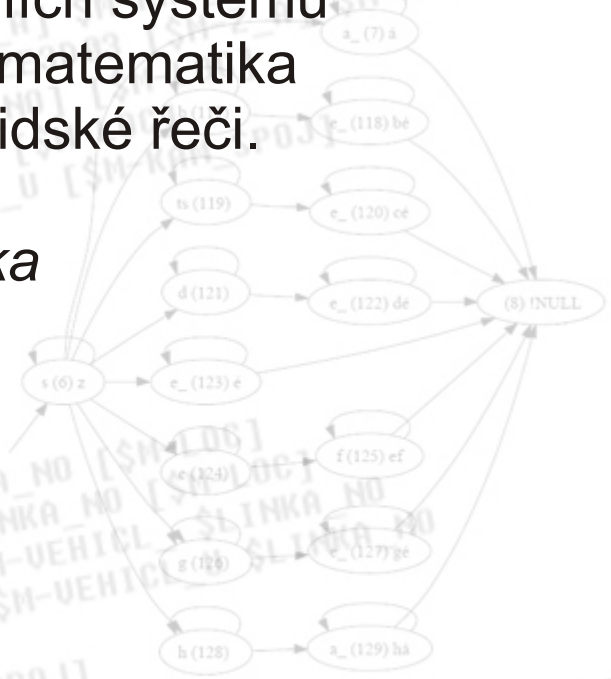
Laboratoř inteligentních komunikačních systémů

Komunikace s počítačem přirozenou řečí

Populárně-naučná přednáška členů výzkumného týmu
Laboratoře inteligentních komunikačních systémů
o tom, že čeština je vlastně vysoká matematika
a jaké to je učit počítače rozumět lidské řeči.

Ekštein | *Konopík* | *Pavelka*



Laboratoř inteligentních komunikačních systémů

Co je to komunikace přirozenou řečí?

- docela obyčejné **povídání si**
- nejběžnější a nejpřirozenější způsob výměny informací mezi lidskými bytostmi
  → **lidská řeč**
- tento způsob je nám vlastní, nezdá se nám složitý, nebráníme se mu → **je přirozený**

O co se v LIKS snažíme?

- abychom mohli lidskou řeč využívat i při komunikaci se stroji, pro začátek s počítači...
- vymýšlíme a vyvíjíme programové vybavení, které umožňuje počítači "poslouchat" uživatele a vykonávat jeho příkazy a odpovídat na jeho dotazy

Proč je užitečné naučit počítač rozumět lidské řeči?

- **usnadnění přístupu k informacím tělesně a smyslově postiženým** a pomoc při jejich integraci do společnosti (ovládání PC bez klávesnice, překlad do znakové řeči, automatické titulkování)
- **nahrazení lidského operátora** v nudných, zdlouhavých, namáhavých a příliš rutinních úlohách (call centra, telefonní rezervace, objednávky, zapisovatelé u soudů, apod.)
- **medicínské aplikace**, zejména v oblasti psychiatrie, klinické psychologie a otorinolaryngologie (rehabilitace afázií, automatická diagnostika psychického stavu, automatická diagnostika dysfonie)
- různé **vojenské a bezpečnostní aplikace** (uvolnění rukou operátora vojenských strojů, monitorování telefonní sítě, špionáž) - *odtud plyne na výzkum nejvíce prostředků*

Co to z hlediska počítače znamená rozumět lidské řeči?

Komunikace přirozenou řečí

Rozpoznávání řeči
(Speech Recognition)

- počítač dokáže převést zvuk zachycený mikrofónem na text v daném jazyce (ale vůbec mu nerozumí)

Porozumění přirozené řeči
(Natural Language Understanding)

- počítač umí (částečně) pochopit, **co** uživatel říká – umí zjistit význam dané posloupnosti slov
- díky tomu pak může uživateli odpovědět nebo jinak správně zareagovat

Co to z hlediska vědce znamená naučit počítač rozumět lidské řeči?

Rozpoznávání řeči
(Speech Recognition)

- počítač dokáže převést zvuk zachycený mikrofónem na text v daném jazyce (ale vůbec mu nerozumí)

Porozumění přirozené řeči
(Natural Language Understanding)

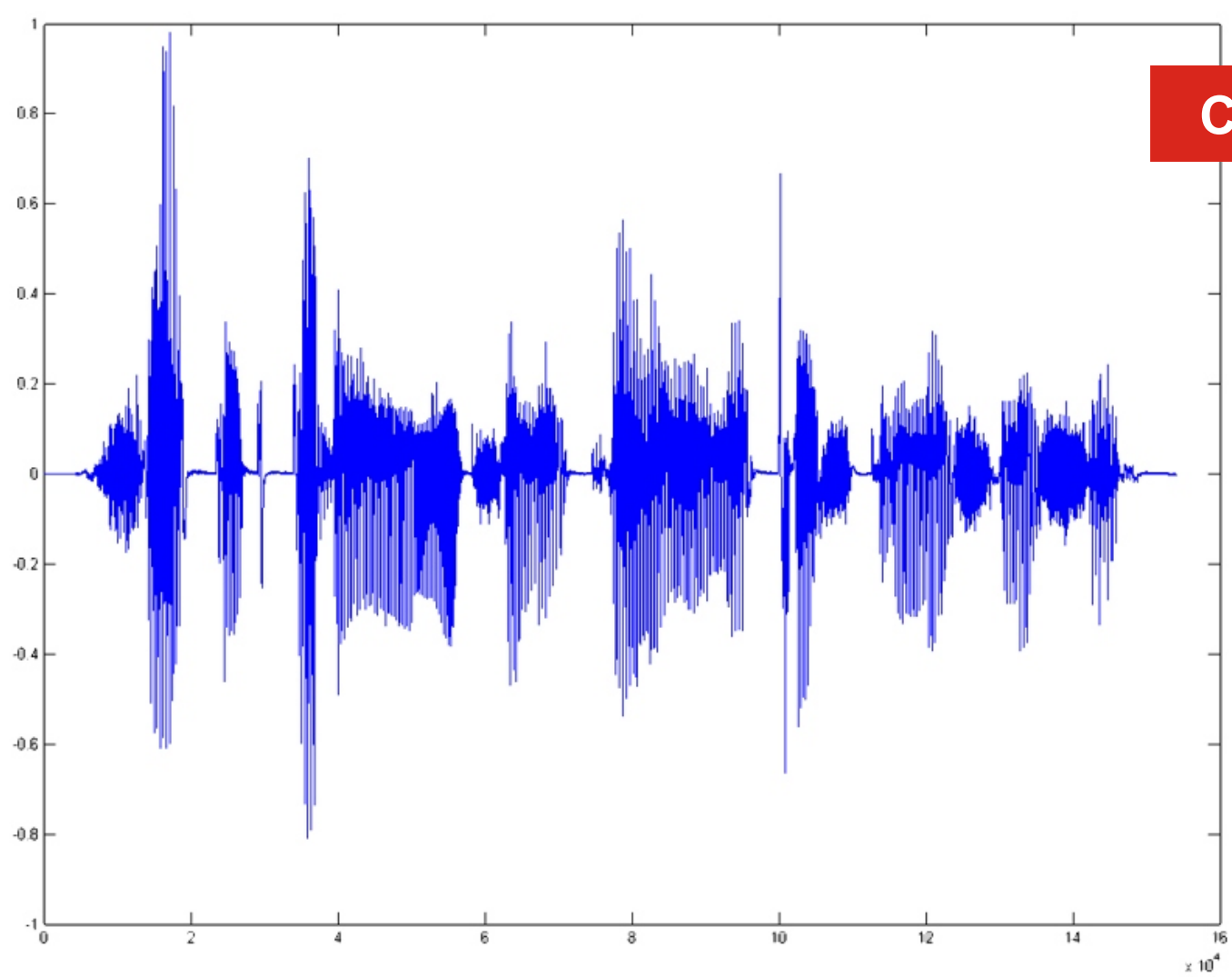
- počítač umí (částečně) pochopit, **co** uživatel říká – umí zjistit význam dané posloupnosti slov
- díky tomu pak může uživateli odpovědět nebo jinak správně zareagovat

Jak to vidí počítač?

0	-0.0002	0.0001	0.0006	0.0004	-0.0001
0	-0.0010	-0.0005	0.0004	0.0006	0.0006
0	-0.0017	-0.0009	-0.0007	-0.0001	0.00
-0.3052	-0.0020	-0.0016	-0.0010	-0.0006	-0.00
0	-0.0016	-0.0009	-0.0007	-0.0008	-0.00
0	-0.0011	-0.0007	-0.0009	-0.0009	-0.0013
0	-0.0007	-0.0001	-0.0007	-0.0002	-0.0013
0.3052	-0.0001	0.0004	0.0004	-0.0003	-0.0010
0	-0.0002	0.0004	0.0011	0.0006	-0.0003
0	-0.0006	-0.0003	0.0003	0.0008	-0.0001
0	-0.0013	-0.0006	-0.0005	-0.0002	0.0003
0	-0.0013	-0.0007	-0.0012	-0.0008	0.0002
-0.3052	-0.0014	-0.0006	-0.0008	-0.0012	-0.0013
0	-0.0010	-0.0006	-0.0008	-0.0009	-0.0016
0	-0.0009	0	-0.0003	-0.0003	-0.0008
0.3052	-0.0006	0.0008	-0.0002	-0.0005	-0.0008
0	-0.0003	0.0011	0.0004	0.0003	-0.0005
-0.3052	-0.0002	0.0004	0.0001	0.0007	0.0002
0	-0.0010	0.0001	-0.0006	0.0005	0.0005
0.3052	-0.0020	-0.0007	-0.0011	-0.0006	0.0002
0	-0.0017	-0.0011	-0.0012	-0.0009	-0.0005
0	-0.0015	-0.0009	-0.0007	-0.0009	-0.0007
0	-0.0009	-0.0009	-0.0004	-0.0006	-0.0011
0	-0.0003	-0.0002	-0.0003	-0.0003	-0.0010
-0.3052	0.0001	0.0006	0.0005	0.0001	-0.0004
0	0.0004	0.0001	0.0009	0.0012	-0.0011
0	-0.0006	-0.0008	0.0001	0.0006	0.0003
0	-0.0014	-0.0010	-0.0009	-0.0001	0
0	-0.0017	-0.0010	-0.0012	-0.0010	-0.0014
0	-0.0009	-0.0008	-0.0008	-0.0007	-0.0013
0	-0.0005	-0.0007	-0.0006	-0.0008	-0.0017

Co to je?

Jak to vidí počítač?

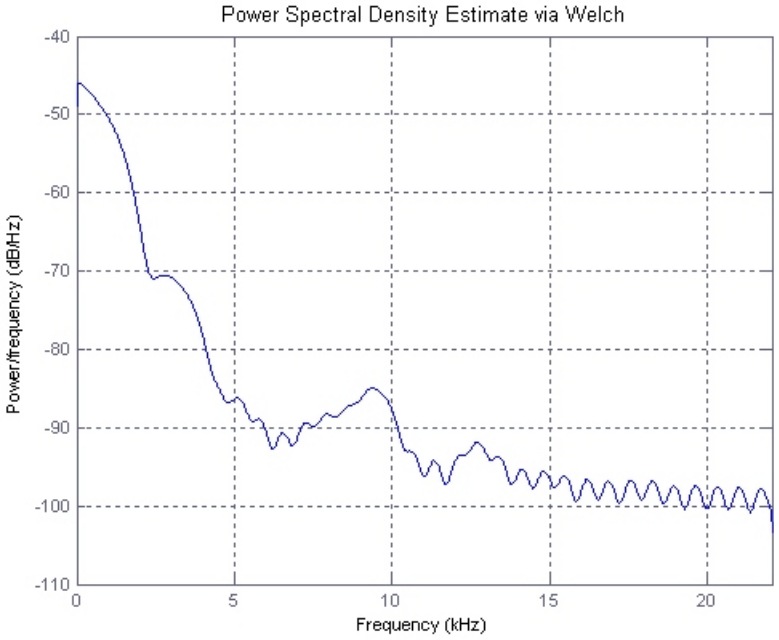


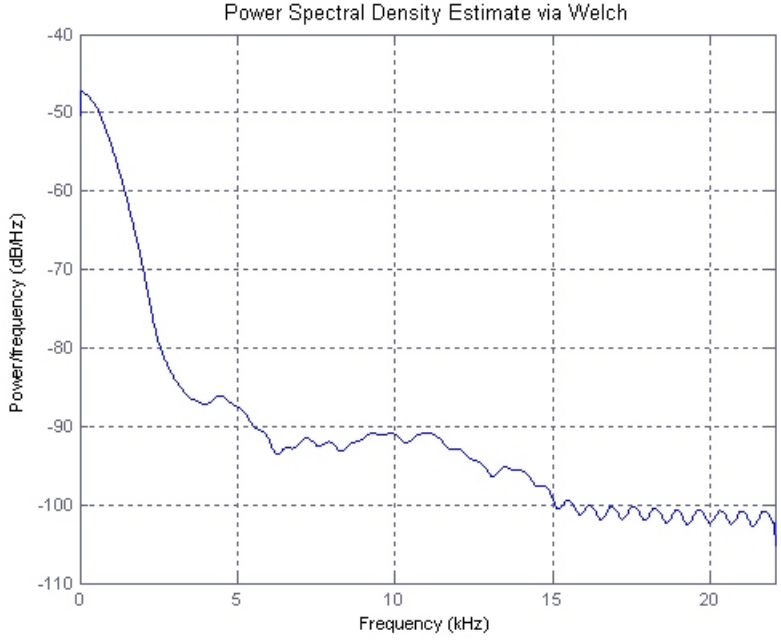
Co to je?

Proč naivní přístupy nefungují?

[ʌ] ve slově **šlapat** [ʃlʌpʌt]

IPA





Realizace stejného fonému se od sebe mohou značně lišit, záleží na kontextu, ve kterém byl foném vysloven, na stavu mluvčího, aj.

Proč naivní přístupy nefungují?

Neexistuje fungující postup, jak porovnat nějaký vzorový signál s pozorovaným signálem → neumíme stanovit **metriku**... (možná spíš taková metrika vůbec nelze stanovit)

Phoneme	ASD/SS	MinSD	MaxSD	Phoneme	ASD/SS	MinSD	MaxSD
[a]	5.0310	2.2930	9.7246	[aː]	5.5524	3.1457	8.5669
[ɛ]	4.1480	2.2303	8.6873	[ɛː]	5.1789	2.8320	7.8600
[ɪ]	4.6924	2.6014	8.6767	[ɪː]	4.0416	2.0703	7.7942
[ɔ]	3.3727	1.4181	9.4446	[ɔː]	3.5215	1.5039	9.1579
[ʊ]	3.1089	1.3521	7.8226	[ʊː]	5.6246	2.9559	8.8400
[p]	3.5096	1.1112	10.0436	[b]	2.8249	1.0169	12.0169
[t]	5.3552	1.6458	8.3022	[d]	3.8286	1.4120	9.3461
[c]	8.8286	3.2104	12.7894	[ɟ]	4.3839	1.9448	7.3808
[k]	5.5369	1.5845	7.9572	[g]	4.4577	2.5143	7.3216
[f]	7.8481	2.9659	12.3889	[v]	4		
[s]	7.3093	3.7270	11.2596	[z]	6.8366	3.4747	10.5655
[ʃ]	7.3554	4.1233	11.2121	[ʒ]	8.8460	3.8291	12.1522
[x]	7.7645	3.6655	11.3786	[ɦ]	4.2399	2.7461	7.8698
[ř]	7				6.0378	3.4957	8.7062
[r]	5.7786	3.0179	9.1160	[l]	4.0855	2.1800	9.2524
[m]	2.5494	1.2147	10.3283	[n]	2.8817	1.3275	9.8148
[ɲ]	2.9749	1.5441	9.3680				

nejméně stabilní foném

nejstabilnější foném

Co tedy funguje?

Modelování dílčích procesů vnímání a zpracování lidské řeči metodami s **vysokou schopností generalizace a tolerance** založenými na **statistice**.

Tato složitá definice říká, že nejprve vysvětlíme počítači (tedy vytvoříme algoritmus), jak si spojit nějaký jev ve vstupních datech s tím, co po něm chceme jako data výstupní. Pak mu předložíme velké množství vzorků vstupních dat spolu s požadovanými daty výstupními – tento proces se nazývá **trénování**.

Jak se pořizují trénovací data?

Do korpusu trénovacích dat se zařazují i nahrávky spontánních mluvených projevů – ty se musí ručně **transkribovat**.
K řečovému signálu se pořídí jeho přesný přepis, včetně zachycení různých zvukových projevů mluvčího (kašlání, nádechy, koktání, aj.).

Na přesnosti transkripce závisí úspěšnost trénování.

5 minut záznamu = 1 hodina transkripce

Studentka Filosofické fakulty Gabriela Wagnerová při transkripci korpusového materiálu.

Jak vypadá transkripce?

transkripční značky

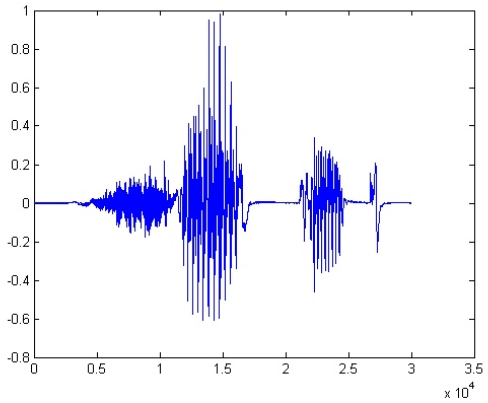
```
+sl +rs no á to +eh co je tady na těch slajdech tak +sl tady
je to demonstrovaný na překladači oupn vatkom +sl cé plus
plus ve verzi jedna pět a +is upozorňuju tedá že se to
samozřejmě může v závislosti na použitym nástroji +eh
výrazně lišit ó- v majkrosoftim céčku to asi bude o dost
jiný +sl +is

ale tři osum šest +eh to už se nezměnilo takže pentium
vlastně využívá pořád tuhletu instrukční sadu s tim že se
tam sice vobjevily ňáký pár novejch instrukcí zejména v
oblasti +is +eh těch +eh tý podpory multimédií jako je
((tjako)) sou ty sady em em ix a es es é- a +eh ((áeim))
dé tři dé nau a podobně +sl ale +is to už nás netrápí
```

neznámé slovo

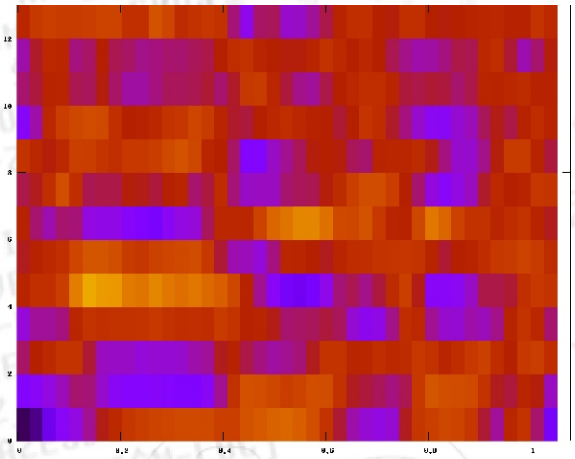
V zápisu musí být uvedené takové tvary slov, jaké mluvčí použil, tj. včetně všech chyb, nesprávné výslovnosti, nespisovné výrazy, atd.

Jak vypadá celý proces rozpoznávání?



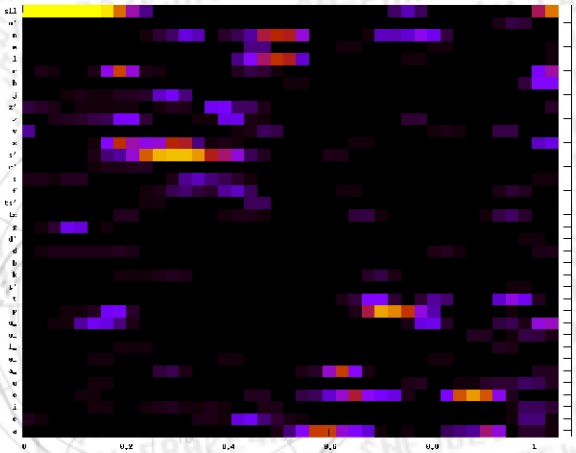
spektrální analýza

(aplikace filtrů a integrálních transformací - FFT, DCT, aj.)



neuronová síť (MLP)

(síť je natrénovaná tak, aby odhadovala, jaký foném představuje dané spektrum)

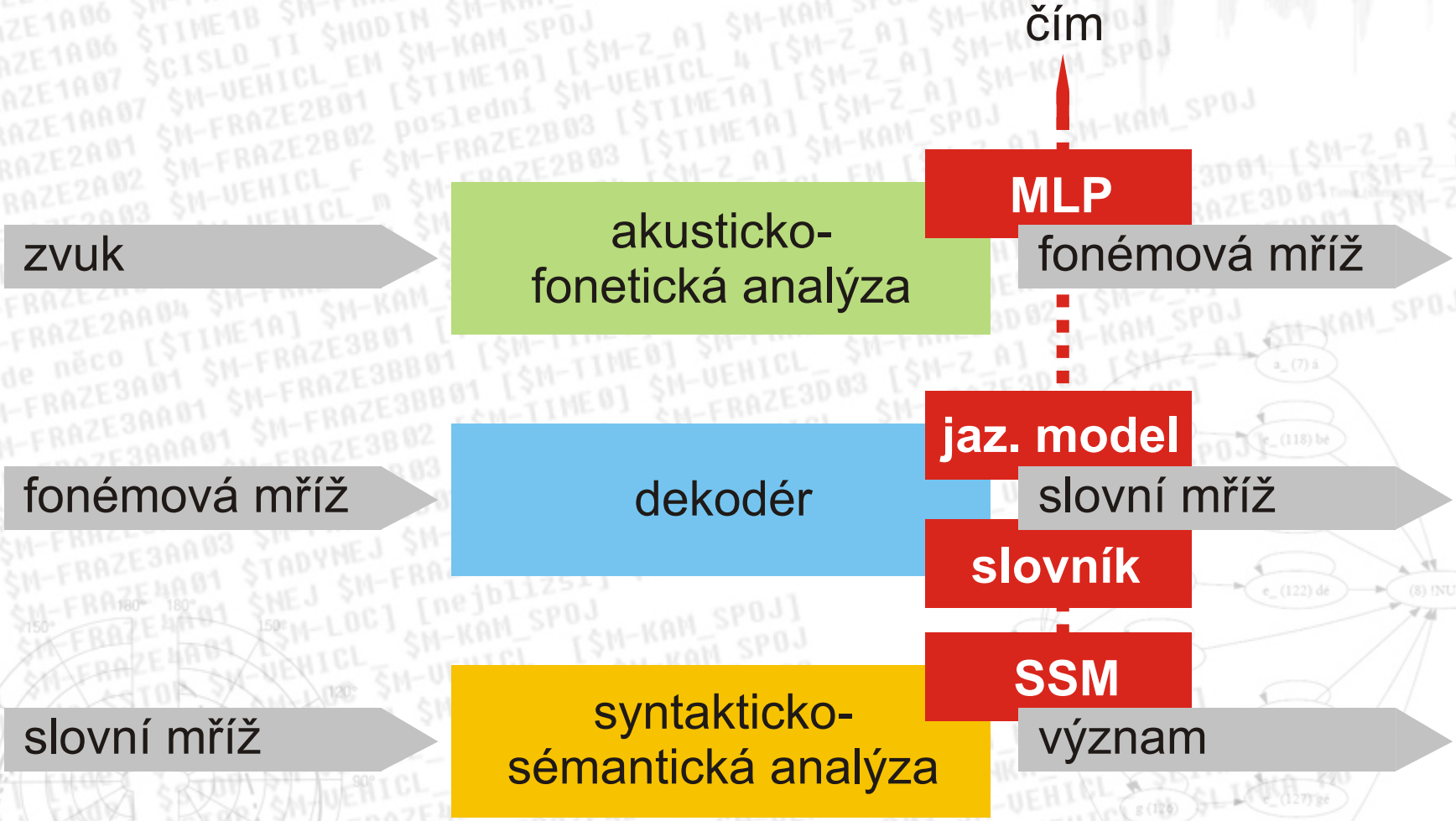


dekodér

(skrytý markovský model)

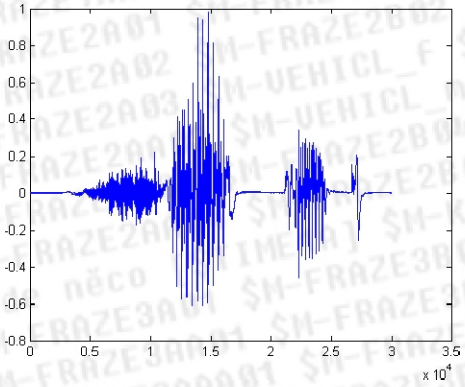
`s' l a p a t`

Jaké procesy se tedy modelují?

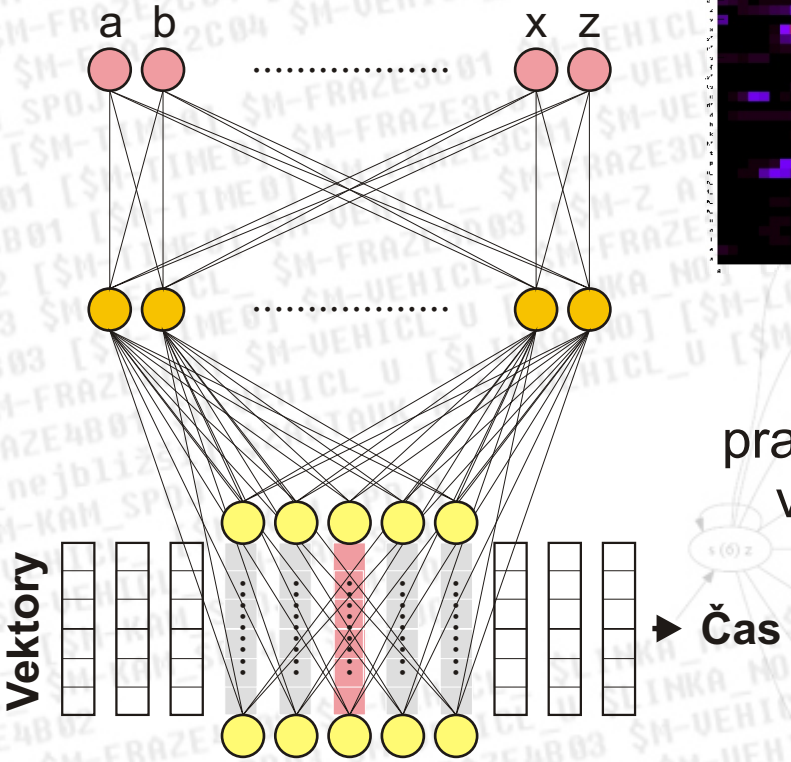


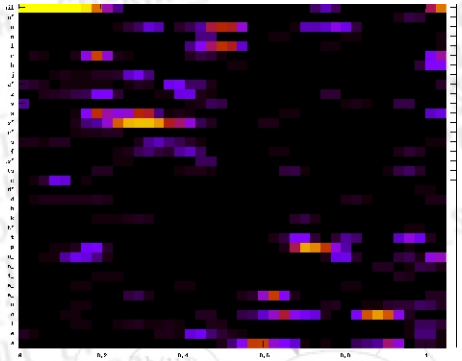
Co je akusticko-fonetická analýza?

zvuk → akusticko-fonetická analýza → fonémová mříž



zvuk je už pře-transformován do podoby tzv. *parametrických vektorů*





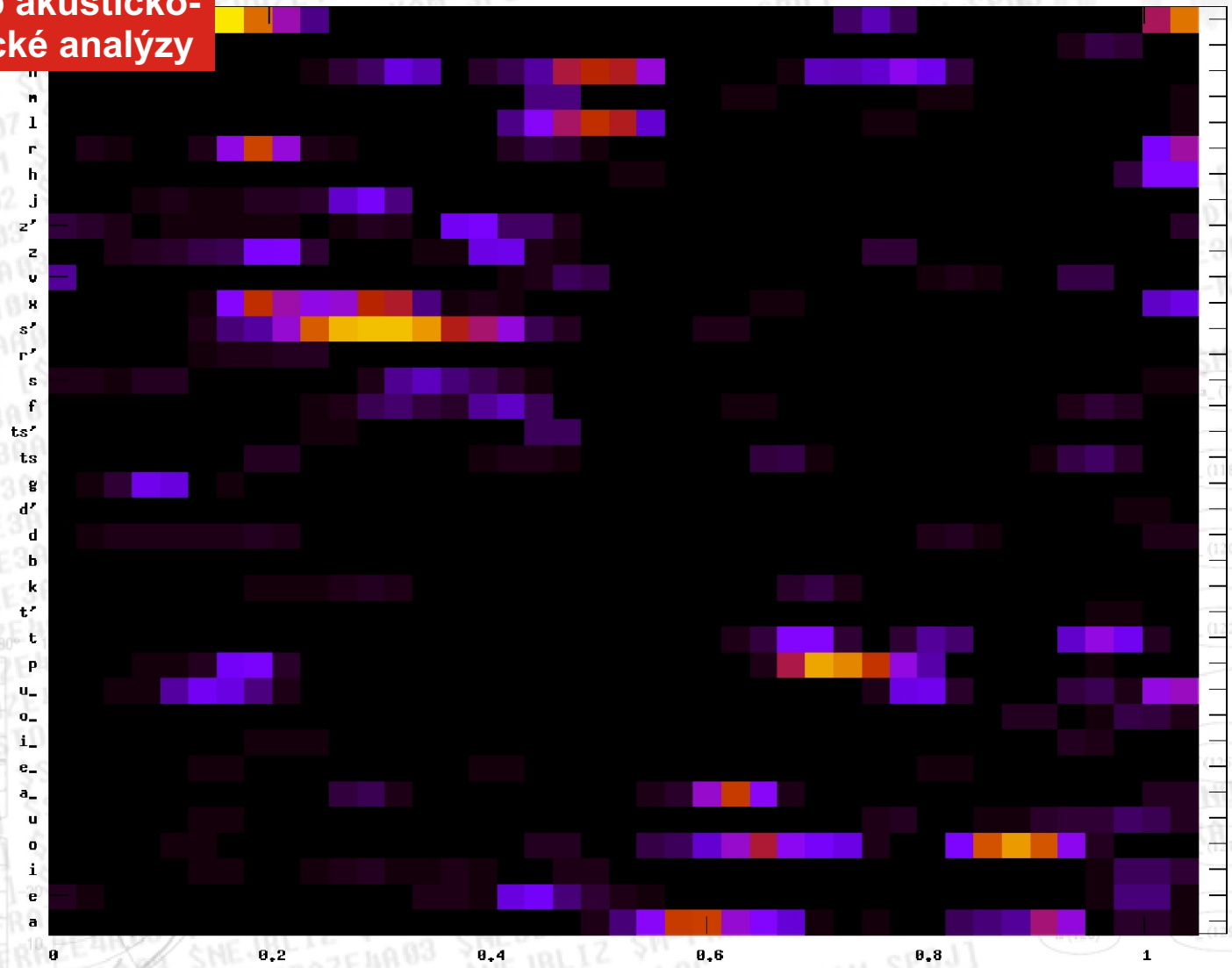
mříž obsahuje pravděpodobnosti výskytu fonému v daném čase

umělá neuronová síť (MLP)

Laboratoř inteligentních komunikačních systémů

Co dělá akusticko-fonetická analýza?

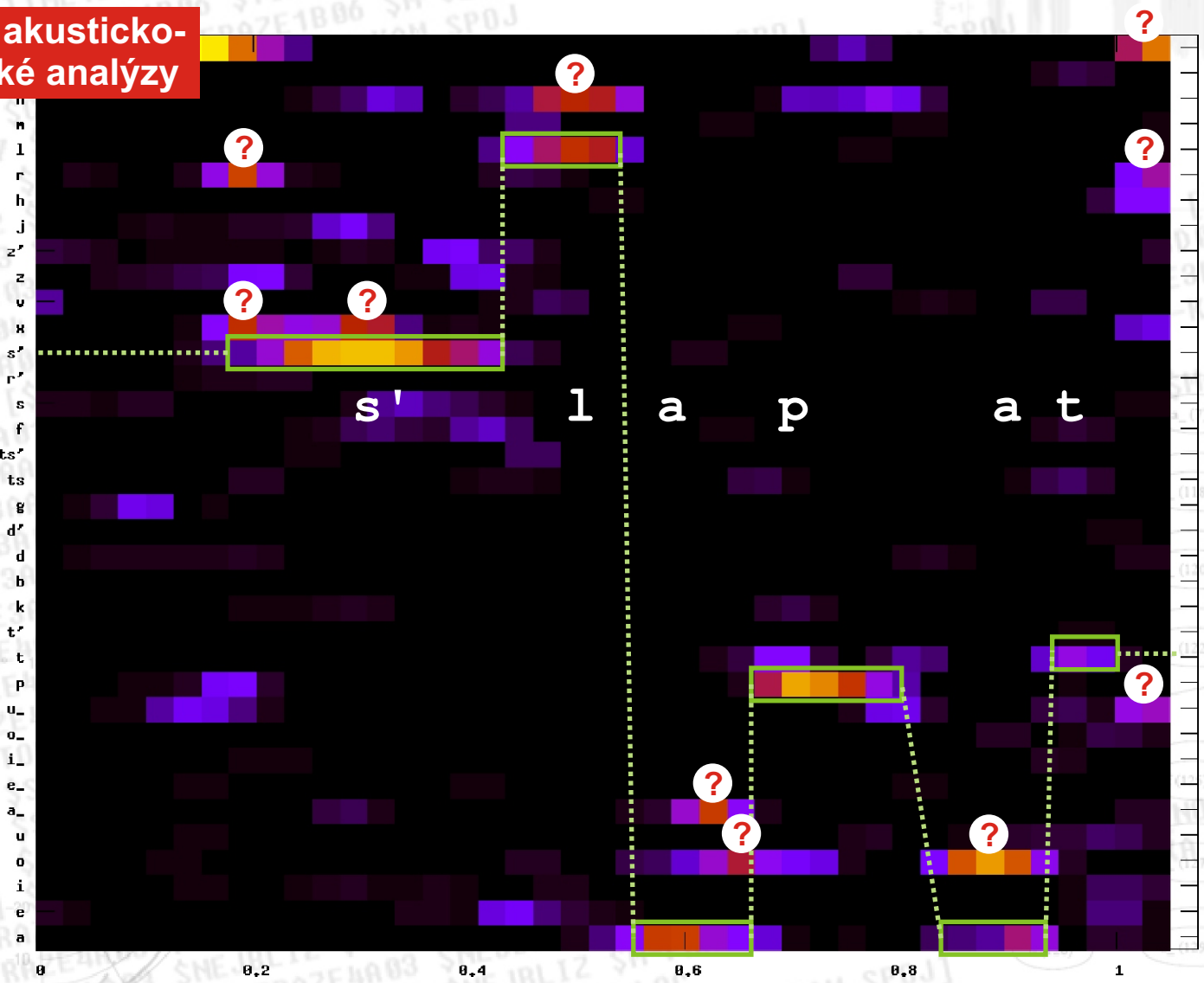
výstup akusticko-fonetické analýzy



Laboratoř inteligentních komunikačních systémů

Co dělá akusticko-fonetická analýza?

výstup akusticko-fonetické analýzy



Jak úspěšně poznává počítač fonémy?

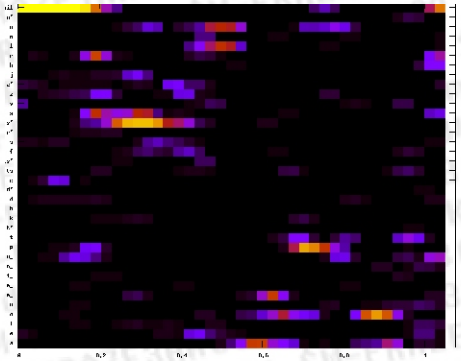
- člověk rozpozná izolovaný foném s úspěšností asi **49%**
  (to jsme vyzkoušeli během zajímavého experimentu)

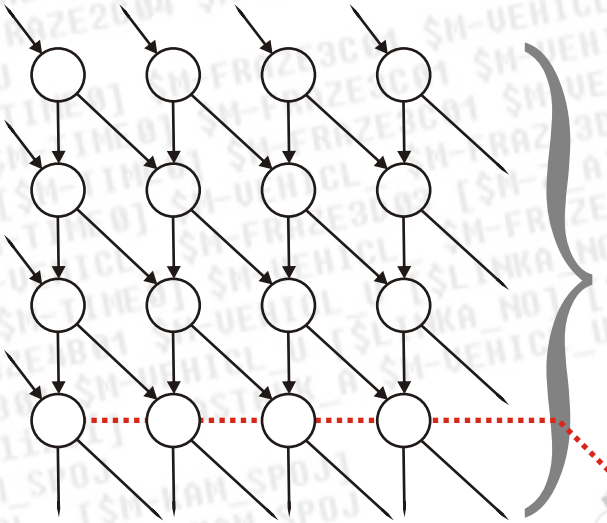
Jaký foném slyšíte? 

- počítač dosahuje při akusticko-fonetické analýze úspěšnosti asi **75%**
- celková úspěšnost rozpoznávání se pohybuje okolo **95 - 97%**
  (slovník má stovky slov, jazykovým modelem je gramatika)
  → úspěšnost je třeba "dohonit" dalšími postupy (jazykovým modelem, pragmatickou analýzou, omezením domény)

Jak nalézt ve fonémové mříži slova?

fonémová mříž → dekodér → slovní mříž





s′	l	a	p	a	t	
s′	l	a		p	a	t

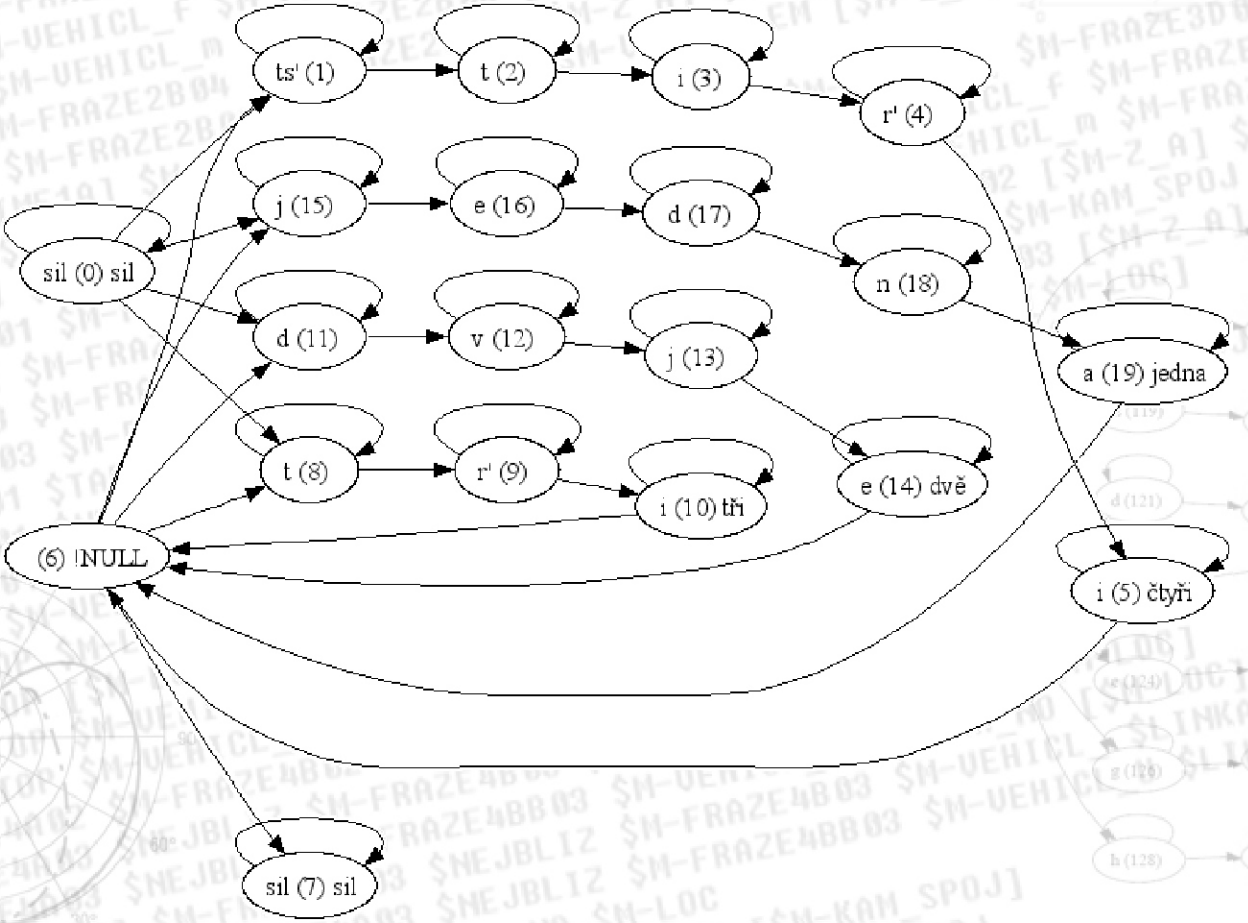
efektivní prohledávání grafu (Viterbiho algoritmus)

stavy modelu

- jako parametr se do dekodéru vloží **graf**, který určuje, které posloupnosti stavy tvoří rozpoznávané promluvy

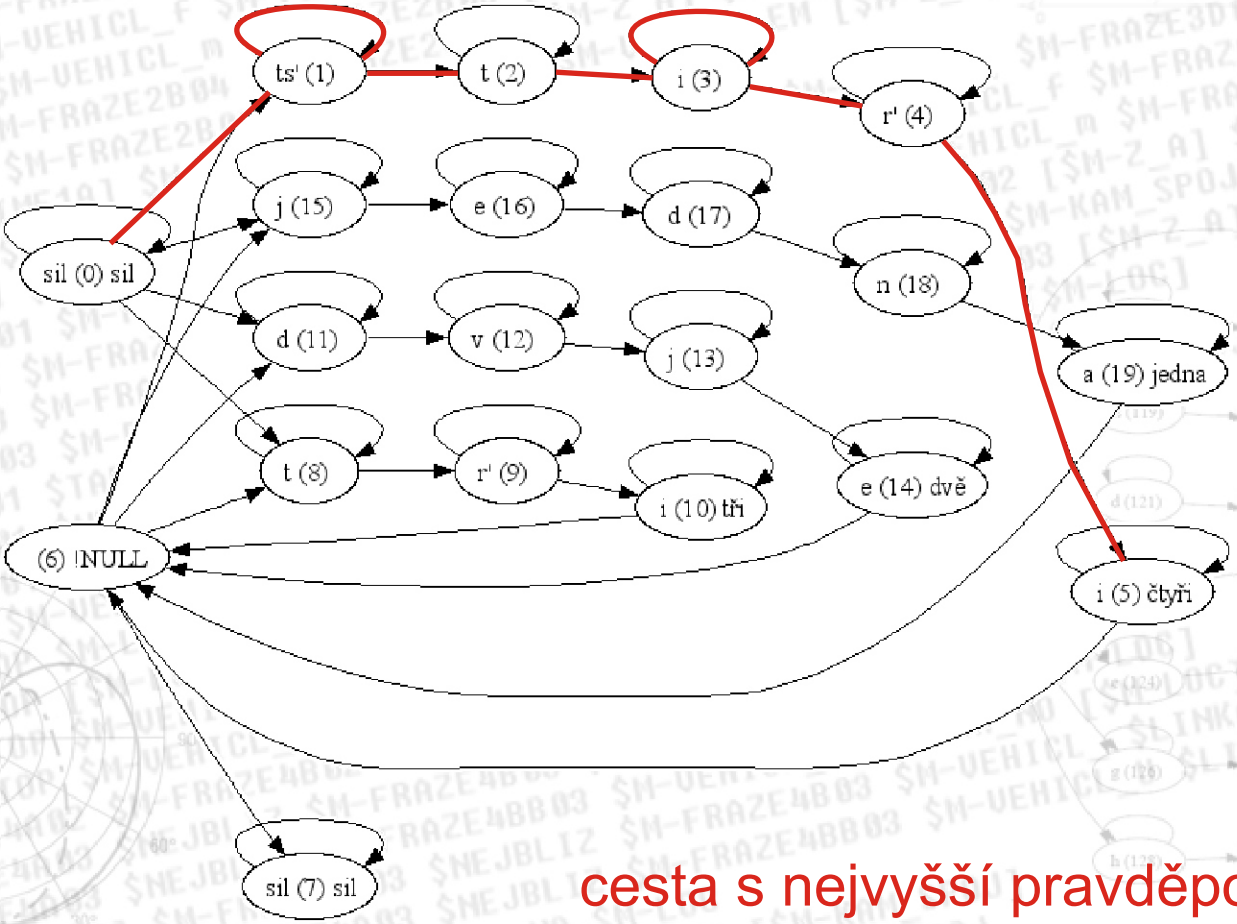
Co dělá dekodér?

- v každém časovém okamžiku je model promluvy v určitém stavu
- graf určuje, do kterých stavů (fonémů) může model přejít
- cílem dekódování je nalezení **nejlepší cesty grafem**



Co dělá dekodér?

- v každém časovém okamžiku je model promluvy v určitém stavu
- graf určuje, do kterých stavů (fonémů) může model přejít
- cílem dekódování je nalezení **nejlepší cesty grafem**



cesta s nejvyšší pravděpodobností

Proč je důležitý jazykový model?

- často se stává, že posloupnost fonémů může tvořit více různých posloupností slov – **většinou nesmyslných**

```
s' l a p a t p o d a_ l n' i ts i n a k …
o l e j e n a p r o s t i_ n e s m i s l
```

A šlapat po dálnici na kole je naprostý nesmysl

B šla pat po dál nic i na koleje na pro stý ne s mysl

?

Co je to jazykový model?

- definuje pravděpodobnosti posloupností slov

jazykové modely

deterministické — **gramatiky**

pravděpodobnost posloupnosti je buď 1 nebo 0 - buď může nastat nebo nemůže

→ **model obvykle vytváří expert**

stochastické — **N-gramy**

pravděpodobnost posloupnosti je v intervalu <0, 1>

→ **model vytváří počítač trénováním z velkého množství trénovacích dat, tzv. korpusu**

Jak vypadá gramatika?

```
$cislice = jedna | dva | tři | čtyři | pět | šest | sedm |
  osm | devět ;
$nact = deset | jedenáct | dvanáct | třináct | čtrnáct | patnáct |
  šestnáct | sedmnáct | osmnáct | devatenáct ;
$desitky = dvacet | třicet | čtyřicet | padesát | šedesát |
  sedmdesát | osmdesát | devadesát ;
$sta = sto | dvě stě | ((tři | čtyři) sta) | ((pět | šest | sedm |
  osm | devět) set) ;
$tisice = (dva | tři | čtyři) tisíce | [jeden | pět | šest | sedm |
  osm | devět | $nact | ($desitky [$cislice]) | ($sta [$nact |
  ($desitky [$cislice])])] tisíc ;
$miliony = [jeden] milion | ((dva | tři | čtyři) miliony) | (pět |
  šest | sedm | osm | devět | $nact | ($desitky [$cislice]) | ($sta
  [$nact | ($desitky [$cislice])])) milionů ;
$miliardy = [jedna] miliarda;

$cis = $cislice | $nact | $desitky [$cislice] | $sta ([$nact] |
  [$desitky] [$cislice]) | $tisice ([$sta] ([$nact] | [$desitky]
  [$cislice])) | $miliony ([$tisice] ([$sta] ([$nact] | [$desitky]
  [$cislice]))) | $miliardy ;

(sil $cis sil)
```

rozšířená Backus-Naurova forma (EBNF)

Dokáže gramatika popsat řeč?

Bohužel ne. Přirozená řeč je tak komplikovaná, že její popis formální gramatikou není proveditelný. I kdyby se někomu podařilo formální gramatiku pro češtinu sestrojit, bude tak obrovská, že ji žádný počítač nebude schopen zpracovat...

Případ českého slovesa:

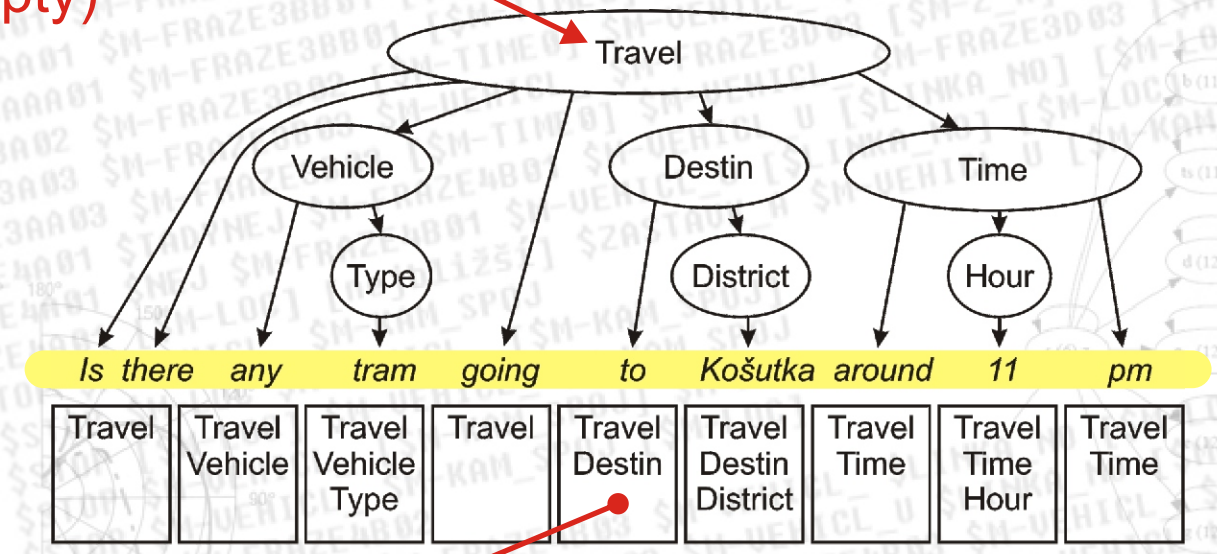
- 6 osob
- 2 rody
- 3 časy
- 3 způsoby
- 2 vidy
- 6 přechodníků

} tvoření vesměs není kompozitní → přibližně **200 různých tvarů**

Co je to sémantika?

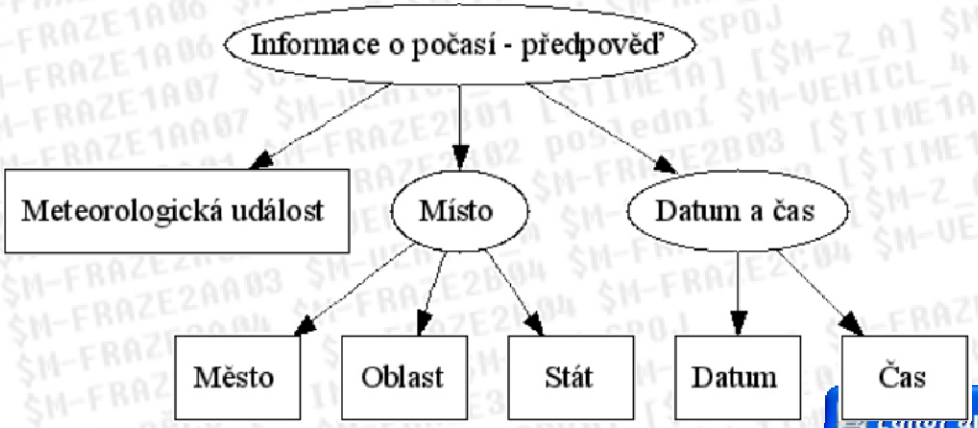
- dobývání významu z pronesené výpovědi
- **SSM** (Stochastic Semantic Model) – gramatika obohacená o pravděpodobnosti příslušnosti výrazu do sémantické kategorie

sémantické kategorie (koncepty)



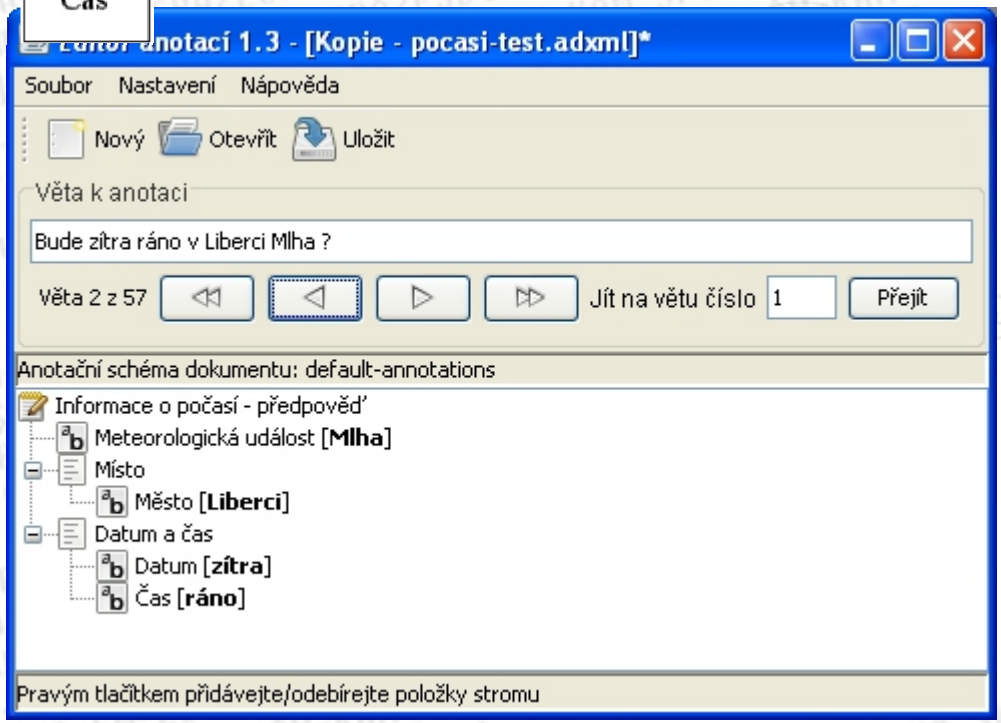
sémantické vektory

Jak se vyrábí sémantický korpus?



Bude zítra ráno v Liberci mlha?

- do korpusu se zařadí velké množství vět (21 000)
- v každé větě je třeba každé slov či spojení zařadit do **sémantické kategorie**



Kde se tohle naučím?

Katedra informatiky a výpočetní techniky
Fakulta aplikovaných věd Západočeské univerzity v Plzni

- bakalářský a navazující magisterský studijní program **Inženýrská informatika**
- studijní program Inteligentní počítačové systémy

http://www.kiv.zcu.cz
http://liks.fav.zcu.cz